Relative Information Loss – An Introduction

Bernhard C. Geiger*, Gernot Kubin*
*Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria
{geiger,gernot.kubin}@tugraz.at

***Abstract*—We introduce a relative variant of information loss to characterize the behavior of deterministic input-output systems. We show that the relative loss is closely related to Rényi's information dimension. We provide an upper bound for continuous input random variables and an exact result for a class of functions (comprising quantizers) with infinite absolute information loss. A connection between relative information loss and reconstruction error is investigated.**

I. Introduction

System theory provides a vast literature of mathematical descriptions of deterministic input-output systems. The gain of a linear system at a specific input frequency is specified by its transfer function, and for the distortion introduced by non-linear components certain single-letter measures (e.g., signal-to-distortion ratio) have been defined. These and the measures introduced for the design of systems (e.g., the mean-squared error) give ample choice to the engineer to characterize a system at hand. However, most of the available descriptions are energy-centered or consider second-order statistics only. A big exception are descriptions of chaotic, autonomous dynamical systems [1].

Recently, however, we observe a trend to employ information-theoretic descriptions and cost functions, especially in machine learning and nonlinear adaptive systems [2]. We believe that system theory would also benefit from single-letter information-theoretic characterizations of deterministic input-output systems, and thus have introduced *information loss* as a possible candidate in [3]. In this work we complement the notion of absolute information loss with its relative version, in order to provide a meaningful measure in cases where the absolute information loss is infinite.

Relative information loss for static functions, or *fractional* information loss, has already been introduced by Watanabe [4] in the context of stationary stochastic processes on finite alphabets. It is also worth mentioning that a rather similar quantity has been used in [5], denoted as *information gain ratio*:

$$\frac{I(C;A)}{H(A)} \tag{1}$$

There, A is an attribute with a finite set of values, C is a class variable, and the value of A for which this measure achieves its maximum is assumed to be the most appropriate root of a decision tree used for classification. In this work we will consider the quantity

$$1-\frac{I(C;A)}{H(A)}=\frac{H(A|C)}{H(A)} \tag{2}$$

and extend its definition to a larger class of random variables.

The paper is organized as follows: We define relative information loss in Section II and analyze its elementary properties in Section III. Section IV is devoted to a class of deterministic systems for which the absolute loss was shown to be infinite. We present a bound for the probability of a reconstruction error in Section V and conclude with a few examples in Section VI.

II. A Definition of Relative Information Loss

We start with recalling the definition given in [3], where the absolute information loss induced by transforming an N-dimensional random variable (RV) $\mathbf{X}$ to another N-dimensional RV $\mathbf{Y}$ by a static function $\mathbf{g}$: $\mathcal{X}\to\mathcal{Y}$, $\mathcal{X},\mathcal{Y}\subseteq\mathbb{R}^N$ was given as

$$L(\mathbf{X}\to\mathbf{Y})=\sup_{\mathcal{P}}\Big(I(\hat{\mathbf{X}};\mathbf{X})-I(\hat{\mathbf{X}};\mathbf{Y})\Big)=H(\mathbf{X}|\mathbf{Y}) \tag{3}$$

where the supremum is over all partitions $\mathcal{P}$ of $\mathcal{X}$, and where $\hat{\mathbf{X}}$ is obtained by quantizing $\mathbf{X}$ according to the partition $\mathcal{P}$ (see Fig. 1).

It was shown in [3], that there exist functions which loose an infinite amount of information; in particular, if the probability measure $P_{\mathbf{X}}$ is absolutely continuous w.r.t. the N-dimensional Lebesgue measure ($P_{\mathbf{X}}\ll\mu^N$), quantizers, limiters, and mappings to subspaces of lower dimensionality suffer from infinite information loss. Since some of these functions also transfer an infinite amount of information (i.e., $I(\mathbf{X};\mathbf{Y})=\infty$), information loss alone obviously does not suffice to fully characterize the function $\mathbf{g}$ in information-theoretic terms.

Thus, we complement this absolute quantity of information loss by a relative one, indicating the percentage of information lost in the function:

Definition 1. Let $\mathbf{X}$ be an N-dimensional RV on the sample space $\mathcal{X}$, and let $\mathbf{Y}$ be obtained by transforming $\mathbf{X}$ with a static function $\mathbf{g}$. We define the *relative information loss* induced by this transform as

$$l(\mathbf{X}\to\mathbf{Y})=\lim_{n\to\infty}\frac{H(\hat{\mathbf{X}}_n|\mathbf{Y})}{H(\hat{\mathbf{X}}_n)} \tag{4}$$

where $\hat{\mathbf{X}}_n=\frac{\lfloor n\mathbf{X}\rfloor}{n}$ (elementwise). The quantity on the left is defined if the limit on the right-hand side exists.

One can consider $\hat{\mathbf{X}}_n$ as being obtained by a vector quantization of $\mathbf{X}$ with quantization bins equal to N-dimensional hypercubes of side length $\frac{1}{n}$ (i.e., using a uniform partition

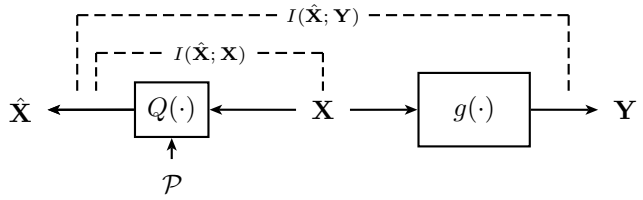

Fig. 1. Model for computing the information loss of a memoryless input-output system g. Q is a quantizer with partition $\mathcal{P}$.

$\mathcal{P}_n$). Note that the partition $\mathcal{P}_{2^{k+1}}$ is a refinement of $\mathcal{P}_{2^k}$ ($\mathcal{P}_{2^{k+1}} \prec \mathcal{P}_{2^k}$).

Remark: First of all, the limit of a sequence of increasingly fine quantizations now takes the place of the supremum in (3). (In Definition 1, the supremum would lead to $l(\mathbf{X} \to \mathbf{Y}) = 1$.) Alternatively, as it was shown in [3], the limit of this sequence can also be used in the Definition of absolute information loss, i.e.,

$$L(\mathbf{X} \to \mathbf{Y}) = \lim_{n\to\infty} H(\hat{\mathbf{X}}_n|\mathbf{Y}) = H(\mathbf{X}|\mathbf{Y}). \tag{5}$$

Again, this only holds if the limit exists.

III. Elementary Properties of Relative Information Loss

We will now highlight the basic properties of relative information loss: First of all, $l(\mathbf{X} \to \mathbf{Y}) \in [0, 1]$, which is due to the non-negativity of entropy and the fact that conditioning reduces entropy. It is interesting to note, however, that $l(\mathbf{X} \to \mathbf{Y}) = 0$ does not imply that the function $\mathbf{g}$ is *information lossless*, i.e., that $L(\mathbf{X} \to \mathbf{Y}) = 0$. While this holds for discrete RVs $\mathbf{X}$ (with finite entropy $H(\mathbf{X})$), for RVs with a continuous component this only means that the absolute information loss is finite. Conversely, $l(\mathbf{X} \to \mathbf{Y}) = 1$ does not imply that the information transfer $I(\mathbf{X};\mathbf{Y})$ is zero. Again, while this holds for discrete RVs, for RVs with a continuous component $l(\mathbf{X} \to \mathbf{Y}) = 1$ implies a finite information transfer. However, we can state the following

Proposition 1. *Let $\mathbf{X}$ be such that $H(\mathbf{X}) = \infty$ and let $l(\mathbf{X} \to \mathbf{Y}) > 0$. Then, $L(\mathbf{X} \to \mathbf{Y}) = H(\mathbf{X}|\mathbf{Y}) = \infty$.*

Proof: We prove the proposition by contradiction. To this end, assume that $H(\mathbf{X}|\mathbf{Y}) = L \leq \infty$. Then,

$$l(\mathbf{X} \to \mathbf{Y}) = \lim_{n\to\infty} \frac{H(\hat{\mathbf{X}}_n|\mathbf{Y})}{H(\hat{\mathbf{X}}_n)} = \lim_{n\to\infty} \inf \frac{H(\hat{\mathbf{X}}_n|\mathbf{Y})}{H(\hat{\mathbf{X}}_n)} \tag{6}$$

$$\leq \lim_{n\to\infty} \inf \frac{H(\mathbf{X}|\mathbf{Y})}{H(\hat{\mathbf{X}}_n)} \tag{7}$$

$$= \lim_{n\to\infty} \inf \frac{L}{H(\hat{\mathbf{X}}_n)} = 0 \tag{8}$$

where the inequality is due to data processing. The last equality follows from the fact that at least a subsequence of $H(\hat{\mathbf{X}}_n)$ converges to $H(\mathbf{X})$ (cf. [6], [7]). ∎

Another interesting property of the sequence

$$\frac{H(\hat{\mathbf{X}}_n|\mathbf{Y})}{H(\hat{\mathbf{X}}_n)} \tag{9}$$

is that, while it might be converging (as we will show in some practically relevant cases below), it is neither generally increasing or decreasing. Consider, for example, a function $\mathbf{g}$ which is bijective if restricted to elements of the partition $\{\mathcal{X}_j\}$, but non-injective on its domain (cf. [3]). Thus, for a partition $\mathcal{P}_{n_0} \prec \{\mathcal{X}_j\}$, and an input probability measure $P_{\mathbf{X}} \ll \mu^N$ the sequence in (9) is decreasing for all further refinements. Conversely, let $\mathbf{g}$ be a vector quantizer with partition $\{\mathcal{X}_j\}$ and let $\mathcal{P}_{n_0} = \{\mathcal{X}_j\}$. Here, while $H(\hat{\mathbf{X}}_{n_0}|Y) = 0$ the sequence in (9) increases for all further refinements (cf. Section VI-A).

Definition 1 has an interesting relationship to the ϵ-entropy proposed by Kolmogorov in [7], [8], but an even more tight connection can be made to the *information dimension* proposed by Rényi in [9]. From there, we restate

Lemma 1 (Asymptotic behavior of $H(\hat{\mathbf{X}}_n)$)**.** *Let $\mathbf{X}$ be an RV with existing information dimension $d(\mathbf{X})$ and let $H(\hat{\mathbf{X}}_1) < \infty$. Then, for $n \to \infty$ the entropy of the RV $\hat{\mathbf{X}}_n$ quantized as in Definition 1 behaves as*

$$H(\hat{\mathbf{X}}_n) = d(\mathbf{X}) \log n + h + o(1) \tag{10}$$

where h is the $d(\mathbf{X})$-dimensional entropy of $\mathbf{X}$ (provided it exists).

Proof: See [9] (cf. also [7], [8]). ∎

For an absolutely continuous RV $\mathbf{X}$ we obtain from this Lemma the following

Corollary 1 (Theorems 1 & 4 in [9])**.** *Let $\mathbf{X}$ be an RV with $P_{\mathbf{X}} \ll \mu^N$ and $H(\hat{\mathbf{X}}_1) < \infty$. Then, for $n \to \infty$ the entropy behaves as*

$$H(\hat{\mathbf{X}}_n) = N \log n + h(\mathbf{X}) + o(1) \tag{11}$$

where $h(\cdot)$ is the differential entropy of $\mathbf{X}$ (provided it exists).

In other words, as a first approximation, the entropy of a continuous RV depends on the dimension of its probability measure, and only as a second approximation on the shape of its density. Note that the second and the third term in Lemma 1 can be neglected for large n.

Using these results we now maintain

Theorem 1. *Let $\mathbf{X}$ be an RV with positive information dimension. Then, if $d(\mathbf{X}|\mathbf{Y} = \mathbf{y})$ exists for all $\mathbf{y} \in \mathcal{Y}$, the relative information loss equals*

$$l(\mathbf{X} \to \mathbf{Y}) = \frac{\mathrm{E}_{\mathbf{Y}}\{d(\mathbf{X}|\mathbf{Y} = \mathbf{y})\}}{d(\mathbf{X})} \tag{12}$$

where $\mathrm{E}_{\mathbf{Y}}\{\cdot\}$ denotes the expectation w.r.t. $\mathbf{Y}$.

Proof: For the proof we use the definition of information dimension given in [9],

$$d(\mathbf{X}) = \lim_{n\to\infty} \frac{H(\hat{\mathbf{X}}_n)}{\log n} \tag{13}$$

where by assumption the limit exists. We obtain

$$l(\mathbf{X}\to\mathbf{Y}) = \frac{\mathrm{E}_{\mathbf{Y}}\left\{d(\mathbf{X}|\mathbf{Y}=\mathbf{y})\right\}}{d(\mathbf{X})} \tag{14}$$

$$= \frac{\int_{\mathcal{Y}} \lim_{n\to\infty} \frac{H(\hat{\mathbf{X}}_n|\mathbf{Y}=\mathbf{y})}{\log n} dP_{\mathbf{Y}}(\mathbf{y})}{\lim_{n\to\infty}\frac{H(\hat{\mathbf{X}}_n)}{\log n}} \tag{15}$$

$$\overset{(a)}{=} \frac{\lim_{n\to\infty}\int_{\mathcal{Y}} \frac{H(\hat{\mathbf{X}}_n|\mathbf{Y}=\mathbf{y})}{\log n} dP_{\mathbf{Y}}(\mathbf{y})}{\lim_{n\to\infty}\frac{H(\hat{\mathbf{X}}_n)}{\log n}} \tag{16}$$

$$\overset{(b)}{=} \lim_{n\to\infty} \frac{\int_{\mathcal{Y}} H(\hat{\mathbf{X}}_n|\mathbf{Y}=\mathbf{y}) dP_{\mathbf{Y}}(\mathbf{y})}{H(\hat{\mathbf{X}}_n)} \tag{17}$$

$$= \lim_{n\to\infty} \frac{H(\hat{\mathbf{X}}_n|\mathbf{Y})}{H(\hat{\mathbf{X}}_n)} \tag{18}$$

where in (a) we used Lebesgue's dominated convergence theorem (e.g., [10]) and where (b) results from the fact that, by assumption, the limits in the numerator and denominator exist and are finite. ∎

This tight connection between relative information loss and the ratio of information dimensions leads to a series of interesting insights, as we will show in this and a companion paper [11]. In particular, it will prove useful if the probability measures are absolutely continuous w.r.t. Lebesgue measure, as information and geometric dimension coincide in this case (cf. [9]).

We are now ready to establish an upper bound on the relative information loss in the following

Theorem 2. *Let $\mathbf{X}$ be an RV with a probability measure $P_{\mathbf{X}} \ll \mu^N$ and with $H(\hat{\mathbf{X}}_1) < \infty$. Then, if the quantities on the right exist,*

$$l(\mathbf{X}\to\mathbf{Y}) \le \frac{1}{N}\sum_{i=1}^{N} l(X^{(i)}\to\mathbf{Y}) \le \frac{1}{N}\sum_{i=1}^{N} l(X^{(i)}\to Y^{(i)}) \tag{19}$$

where $X^{(i)}$ and $Y^{(i)}$ are the components of $\mathbf{X}$ and $\mathbf{Y}$, respectively.

Proof: With Definition 1 and the chain rule of entropy we get

$$l(\mathbf{X}\to\mathbf{Y}) = \lim_{n\to\infty} \frac{\sum_{i=1}^{N} H(\hat{X}_n^{(i)}|\hat{X}_n^{(1)},\ldots,\hat{X}_n^{(i-1)},\mathbf{Y})}{H(\hat{\mathbf{X}}_n)}$$

$$\le \lim_{n\to\infty} \frac{\sum_{i=1}^{N} H(\hat{X}_n^{(i)}|\mathbf{Y})}{H(\hat{\mathbf{X}}_n)} \tag{20}$$

$$\overset{(a)}{=} \frac{1}{d(\mathbf{X})}\sum_{i=1}^{N} \mathrm{E}_{\mathbf{Y}}\left\{d(X^{(i)}|\mathbf{Y}=\mathbf{y})\right\} \tag{21}$$

where in (a) we exchanged summation and limit for similar reasons as in the proof of Theorem 1. Corollary 1 now tells us that due to the absolute continuity $d(\mathbf{X}) = N$ and $d(X^{(i)}) = 1$ for all i. We thus obtain

$$l(\mathbf{X}\to\mathbf{Y}) \le \frac{1}{N}\sum_{i=1}^{N} \frac{\mathrm{E}_{\mathbf{Y}}\left\{d(X^{(i)}|\mathbf{Y}=\mathbf{y})\right\}}{d(X^{(i)})} = \frac{1}{N}\sum_{i=1}^{N} l(X^{(i)}\to\mathbf{Y}) \tag{22}$$

which proves the first inequality. The second inequality is obtained by bounding $H(\hat{X}_n^{(i)}|\mathbf{Y}) \le H(\hat{X}_n^{(i)}|Y^{(i)})$ in (20). ∎

At this point it is worth noting that throughout Section III no assumptions about a functional dependence between $\mathbf{X}$ and $\mathbf{Y}$ were made. Indeed, all statements made in this Section hold equally for stochastic relationships (including stochastic independence) between $\mathbf{X}$ and $\mathbf{Y}$.

IV. Relative Information Loss for Functions which are Constant

We now apply the relative information loss of Definition 1 to a class of functions for which we showed in [3] that the absolute information loss is infinite. In particular, we are talking about functions which are constant on subsets A_i of the domain with positive probability measure.

To this end, let $P_{\mathbf{X}} \ll \mu^N$ be concentrated on a compact set $\mathcal{X} \subseteq \mathbb{R}^N$. Let further $A_i \subseteq \mathcal{X}$ with $P_{\mathbf{X}}(A_i) > 0$. Without loss of generality, we assume that the subsets A_i are disjoint. Now take $\mathbf{Y} = \mathbf{g}(\mathbf{X})$, where $\mathbf{g}$: $\mathcal{X}\to\mathcal{Y}$, $\mathcal{Y}\subseteq\mathbb{R}^N$, is surjective, measurable, and constant on A_i, i.e., $\mathbf{g}(A_i) = \mathbf{y}_i$. As a consequence, $P_{\mathbf{Y}}$ is atomic on $\{\mathbf{y}_i\}$ (thus, $L(\mathbf{X}\to\mathbf{Y}) = \infty$; cf. [3, Corollary 2]). With $A = \bigcup_i A_i$ we further require that $\mathbf{g}$ is piecewise bijective[1] on $\mathcal{X}\setminus A$, from which follows that $P_{\mathbf{Y}}$ is absolutely continuous on $\mathcal{Y}\setminus\{\mathbf{y}_i\}$.

We can now state the following

Proposition 2. *Let $\mathbf{X}$ be an RV with probability measure $P_{\mathbf{X}} \ll \mu^N$ concentrated on a compact set $\mathcal{X}\subseteq\mathbb{R}^N$. Let $\mathbf{g}$ be such that it is constant on sets A_i of positive $P_{\mathbf{X}}$-measure and piecewise bijective elsewhere. Then, the relative information loss is*

$$l(\mathbf{X}\to\mathbf{Y}) = P_{\mathbf{X}}(A) \tag{23}$$

where $A = \bigcup_i A_i$.

Proof: By assumption and Corollary 1 we have $d(\mathbf{X}) = N$. Due to the properties of the function $\mathbf{g}$, $P_{\mathbf{Y}}$ decomposes into a component $P_{\mathbf{Y}}^{ac} \ll \mu^N$ and an atomic component $P_{\mathbf{Y}}^{d}$ concentrated on the points $\mathbf{y}_i = \mathbf{g}(A_i)$. Thus, the preimage of points $\mathbf{y}_i$ with positive $P_{\mathbf{Y}}$-measure is the union of the set A_i and a countable number of points $\{x_{i,j}\}$. Since the set A_i has positive $P_{\mathbf{X}}$-measure (otherwise $P_{\mathbf{Y}}(\mathbf{y}_i) = 0$), the conditional probability measure $P_{\mathbf{X}|\mathbf{Y}=\mathbf{y}_i} \ll \mu^N$. Due to the compactness of the support $\mathcal{X}$ the conditional entropy $H(\hat{\mathbf{X}}_1|\mathbf{Y}=\mathbf{y}_i) < \infty$, thus the associated information dimension exists and equals N. For all other points $\mathbf{y}\in\mathcal{Y}\setminus\{\mathbf{y}_i\}$

[1] see [3] for a possible definition

the preimage is a countable union of points. The associate conditional probability measure is 0-dimensional.

We now prove this Proposition with the help of Theorem 1:

$$l(\mathbf{X} \to \mathbf{Y}) = \frac{1}{N} \int_{\mathcal{Y}} d(\mathbf{X}|\mathbf{Y} = \mathbf{y}) dP_{\mathbf{Y}}(\mathbf{y}) \tag{24}$$

$$= \frac{1}{N} \int_{\mathcal{Y} \setminus \{\mathbf{y}_i\}} d(\mathbf{X}|\mathbf{Y} = \mathbf{y}) dP_{\mathbf{Y}}(\mathbf{y}) + \frac{1}{N} \sum_i d(\mathbf{X}|\mathbf{Y} = \mathbf{y}_i) P_{\mathbf{Y}}(\mathbf{y}_i) \tag{25}$$

$$= \sum_i P_{\mathbf{Y}}(\mathbf{y}_i) \tag{26}$$

Since the preimage of $\mathbf{y}_i$ under $\mathbf{g}$ consists of a set A_i of positive $P_{\mathbf{X}}$-measure and (zero-measure) points, we can write

$$l(\mathbf{X} \to \mathbf{Y}) = \sum_i P_{\mathbf{Y}}(\mathbf{y}_i) = \sum_i P_{\mathbf{X}}(A_i) = P_{\mathbf{X}}(A) \tag{27}$$

where the last equality follows from the fact that A_i are disjoint and the additivity of the measure $P_{\mathbf{X}}$. ∎

The interesting implication of this result is that the shape of the PDF on A has no influence on the relative loss, and neither has the number of different sets A_i (with different output values $\mathbf{y}_i$) – yet, all these do have an influence on the information transport $I(\mathbf{X}; \mathbf{Y})$. This is in conflict with intuition, which suggests that whatever influences information transfer should also influence information loss, and, thus, also relative information loss. Yet, both the properties in Section III and the fact that $H(\hat{\mathbf{X}}_n)$, as a first approximation, depends more on the dimension and the quantization bin size than on the shape of the PDF [7] confirm this theoretical result.

Furthermore, in this particular case it turns out that $\mathbf{Y}$ is a mixture of a continuous and a discrete RV with information dimension $1 - P_{\mathbf{X}}(A)$ [9], [12]. One is thus led to the conjecture that indeed under some circumstances one can show that

$$l(\mathbf{X} \to \mathbf{Y}) = 1 - \frac{d(\mathbf{Y})}{d(\mathbf{X})}. \tag{28}$$

If this really holds and under which conditions it does is currently under investigation.

V. Relative Information Loss and Reconstruction Error

We next want to find connections between the relative information loss and the probability of a reconstruction error given by

$$P_e = \min_f \Pr(\mathbf{X} \neq f(\mathbf{Y})) \tag{29}$$

where f is a function that tries to estimate or reconstruct the original $\mathbf{X}$ from its image $\mathbf{Y}$. It is well known that Fano's inequality does not hold for countably infinite alphabets (e.g. [13]). However, we employ Fano's inequality here to derive a relationship between relative information loss and the probability of a reconstruction error by starting from a finite alphabet and then taking the limit. We present

Theorem 3. *Let $\mathbf{X}$ be a RV with a probability measure $P_{\mathbf{X}} \ll \mu^N$ which is concentrated on a compact set $\mathcal{X} \subset \mathbb{R}^N$. Let P_e denote the probability of a reconstruction error. Then, the error probability is bounded by the relative information loss from below, i.e.,*

$$P_e \geq l(\mathbf{X} \to \mathbf{Y}). \tag{30}$$

Proof: For the proof we start with a quantized version of the input RV, $\hat{\mathbf{X}}_n$. Since $\hat{\mathbf{X}}_n$ is a discrete RV on a finite alphabet $\hat{\mathcal{X}}_n$, we can employ the standard Fano bound [14],

$$H(\hat{\mathbf{X}}_n|\mathbf{Y}) \leq H_2(P_{e,n}) + P_{e,n} \log \operatorname{card}(\hat{\mathcal{X}}_n) \tag{31}$$

where

$$P_{e,n} = \Pr(\hat{\mathbf{X}}_n \neq f^*(\mathbf{Y})). \tag{32}$$

Since Fano's inequality holds for arbitrary estimators f^*, we let f^* be the composition of $f^\circ = \arg\min_f \Pr(\mathbf{X} \neq f(\mathbf{Y}))$ and the quantizer of Definition 1. $P_{e,n}$ is the probability that $f^\circ(\mathbf{Y})$ and $\mathbf{X}$ do not lie in the same quantization bin. Since the bin volume reduces with n, $P_{e,n}$ increases monotonically to P_e. We thus obtain with $H_2(p) \leq 1$ for all $0 \leq p \leq 1$

$$H(\hat{\mathbf{X}}_n|\mathbf{Y}) \leq 1 + P_e \log \operatorname{card}(\hat{\mathcal{X}}_n). \tag{33}$$

We next define the *diameter* D of $\mathcal{X}$ as

$$D = \sup_{x_1, x_2 \in \mathcal{X}} ||x_1 - x_2|| \tag{34}$$

where $||\cdot||$ is the Euclidean distance and where $D < \infty$ due to the compactness of $\mathcal{X}$. As an immediate consequence, $\mathcal{X}$ can be covered by an N-dimensional hypercube with side length D. Quantizing $\mathbf{X}$ with a vector quantizer corresponds to covering $\mathcal{X}$ by hypercubes of side length $\frac{1}{n}$. It thus follows that

$$\operatorname{card}(\hat{\mathcal{X}}_n) \leq (\lceil nD \rceil)^N \leq (nD + 1)^N \tag{35}$$

and finally

$$H(\hat{\mathbf{X}}_n|\mathbf{Y}) \leq 1 + P_e N \log(nD + 1). \tag{36}$$

With Corollary 1 we thus get

$$l(\mathbf{X} \to \mathbf{Y}) = \lim_{n \to \infty} \frac{H(\hat{\mathbf{X}}_n|\mathbf{Y})}{H(\hat{\mathbf{X}}_n)} \tag{37}$$

$$\leq \lim_{n \to \infty} \frac{1 + P_e N \log(nD + 1)}{H(\hat{\mathbf{X}}_n)} \tag{38}$$

$$\overset{(a)}{=} \lim_{n \to \infty} \frac{1}{N \log n} + \frac{P_e N \log(nD + 1)}{N \log n} \tag{39}$$

$$= P_e \tag{40}$$

where in (a) we again used Theorem 1 and the fact that $d(\mathbf{X}) = N$. This completes the proof. ∎

VI. Examples

In this Section we will now illustrate the theoretical results at the hand of a few examples.

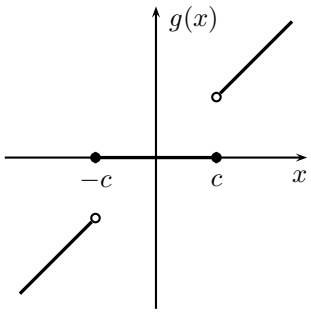

Fig. 2. Center clipper of Example 2

A. Quantizers

The first practical application of our results will be the analysis of a quantizer, which is typically used to represent a continuous RV by a discrete RV, designed according to some optimality criterion (mean-squared reconstruction error, maximum output entropy, etc.). Since the output of the quantizer is discrete in amplitude, it is clear that an infinite amount of information is lost. In addition to that, since the quantizer function is constant almost everywhere it turns out that the relative information loss is unity:

$$l(\mathbf{X} \to \mathbf{Y}) = 1 \tag{41}$$

In other words, disrespective of the (finite) number of quantization bins and the design criterion, the quantizer always destroys 100% of the available information. This holds equally for scalar and vector quantizers. Note, however, that despite this fact still a positive amount of information is transferred by the quantizer (cf. Section III).

B. Center Clipper

In signal processing center clippers (see Fig. 2) are used for noise suppression or residual echo cancellation [15]. We let the center clipper be described by the following function:

$$g(x) = \begin{cases} x, & \text{if } |x| > c \\ 0, & \text{else} \end{cases} \tag{42}$$

By Theorem 2 the relative information loss evaluates to $l(X \to Y) = P_X([-c, c])$, which reveals that it depends only on the clipping parameter c and the probability mass contained in that interval. Yet, since center clippers *do enhance* signal quality in many cases, this suggests that probably a different measure of information loss could be more appropriate.

Note further that the center clipper is bijective if it is restricted to $\mathcal{X} \setminus [-c, c]$. Thus, while outside of $[-c, c]$ we have a zero probability of a reconstruction error, within the center interval the error probability is unity. As a consequence, $P_e = P_X([-c, c])$ which makes the bound of Theorem 3 tight. If g was not bijective outside of $[-c, c]$, but, e.g., would destroy the sign information, then $P_e > P_X([-c, c])$ and Theorem 3 still holds.

VII. Conclusion

In this work, we introduced the notion of relative information loss, complementing its absolute variant presented by the authors in a previous work. We showed that there is a close connection between the relative loss and the Rényi information dimension of the input and the conditional random variable of the input given the output.

For a continuous-valued input both upper bounds and an exact expression for a certain class of systems was presented. In particular, it was shown that quantizers loose 100% of the available information. We finally analyzed a connection between the probability of reconstruction error and relative information loss.

Acknoledgment

The authors thank Yihong Wu, Wharton School, University of Pennsylvania, for bringing Rényi's information dimension to our attention.

Appendix

We now show the following

Lemma 2.

$$\lim_{n\to\infty} H(\hat{\mathbf{X}}_n|\mathbf{Y}) = H(\mathbf{X}|\mathbf{Y}) \tag{43}$$

provided the limit exists.

Proof: For the proof we note that

$$H(\hat{\mathbf{X}}_n|\mathbf{Y}) = I(\mathbf{X}; \hat{\mathbf{X}}_n|\mathbf{Y}) \tag{44}$$

because $\hat{\mathbf{X}}_n$ is a function of $\mathbf{X}$ [6, Ch. 3.9]. Further, if $\xi = (\xi_1, \xi_2, \dots)$ we obtain with [6, Thm. 3.10.1]

$$\lim_{n\to\infty} I((\xi_1, \xi_2, \dots, \xi_n); \eta|\epsilon) = I(\xi; \eta|\epsilon). \tag{45}$$

We now identify $\epsilon = \mathbf{Y}$ and $\eta = \mathbf{X}$. Furthermore, if the limit in Lemma 2 exists, all subsequences converge to the same limit. In particular, also the subsequence $\hat{\mathbf{X}}_{2^k}$ converges to the same limit. We now identify this RV with the binary expansion of $\mathbf{X}$ up to order k; thus, $\hat{\mathbf{X}}_{2^k} = (\xi_1, \xi_2, \dots, \xi_k)$. Clearly, $\lim_{k\to\infty} \hat{\mathbf{X}}_{2^k} = \mathbf{X}$. Comparing this to (45) completes the proof. ∎

References

[1] J. Jost, *Dynamical Systems: Examples of Complex Behavior*. New York, NY: Springer, 2005.

[2] J. C. Principe, *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*, ser. Information Science and Statistics. New York, NY: Springer, 2010.

[3] B. C. Geiger and G. Kubin, "On the information loss in memoryless systems: The multivariate case," in *Proc. Int. Zurich Seminar on Communications (IZS)*, Zurich, Feb. 2012, pp. 32–35, extended version available: `arXiv:1109.4856 [cs.IT]`.

[4] S. Watanabe and C. T. Abraham, "Loss and recovery of information by coarse observation of stochastic chain," *Information and Control*, vol. 3, no. 3, pp. 248–278, Sep. 1960.

[5] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, pp. 81–106, 1986.

[6] M. S. Pinsker, *Information and Information Stability of Random Variables and Processes*. San Francisco, CA: Holden Day, 1964.

[7] A. N. Kolmogorov, "On the Shannon theory of information transmission in case of continuous signals," *IEEE Trans. Inf. Theory*, vol. 2, pp. 102–108, Dec. 1956.

[8] ——, "ϵ-entropy and ϵ-capacity of sets in functional spaces," in *Selected Works of A. N. Kolmogorov – Volume III: Information Theory and the Theory of Algorithms*, A. N. Shiryayev, Ed. Dordrecht: Kluwer, 1993, pp. 86–170.
[9] A. Rényi, "On the dimension and entropy of probability distributions," *Acta Mathematica Hungarica*, vol. 10, pp. 193–215, 1959.
[10] W. Rudin, *Real and Complex Analysis*, 3rd ed. New York, NY: McGraw-Hill, 1987.
[11] B. C. Geiger and G. Kubin, "Relative information loss in the PCA," Apr. 2012, `arXiv:1202.???? [cs.IT]`.
[12] Y. Wu and S. Verdú and, "Rényi information dimension: Fundamental limits of almost lossless analog compression," *IEEE Trans. Inf. Theory*, vol. 56, no. 8, pp. 3721–3748, Aug. 2010.
[13] S.-W. Ho and S. Verdú, "On the interplay between conditional entropy and error probability," *IEEE Trans. Inf. Theory*, vol. 56, no. 12, pp. 5930–5942, Dec. 2010.
[14] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ: Wiley Interscience, 2006.
[15] P. Vary and R. Martin, *Digital speech transmission: Enhancement, coding and error concealment*. Chichester: John Wiley & Sons, 2006.